Typical rank of coin-toss power-law random matrices over $\mathbb{GF}(2)$

Salvatore Mandrà[1,2,3], Marco Cosentino Lagomarsino[1,4,5], and Bruno Bassetti[1,2]

[1]Università degli Studi di Milano, Dip. Fisica, Milano, Italy
[2]INFN, Milano, Italy
[3]Universitat de Barcelona, Dep. Física Fonamental, Barcelona, Spain
[4]Génophysique / Genomic Physics Group, FRE 3214 CNRS "Microorganism Genomics"
[5]University Pierre et Marie Curie, 15, rue de l'École de Médecine 75006 Paris France

Abstract

Random linear systems over the Galois Field modulo 2 have an interest in connection with problems ranging from computational optimization to complex networks. They are often approached using random matrices with Poisson-distributed or finite column/row-sums. This technical note considers the typical rank of random matrices belonging to a specific ensemble wich has genuinely power-law distributed column-sums. For this ensemble, we find a formula for calculating the typical rank in the limit of large matrices as a function of the power-law exponent and the shape of the matrix, and characterize its behavior through "phase diagrams" with varying model parameters.

1 Introduction

This technical note presents the calculation of the typical rank of Boolean random matrices with power-law distributed column-sums. The specificity of this calculation is that it applies to genuinely power-law matrices, without finite cutoffs in the distribution. Before presenting the results, we will give a brief description of the context that motivates the calculation.

Random matrices with Boolean entries are often simple to treat, which makes them important in many paradigmatic problems of different branches of science. For example, in computer science, they define the so-called random XOR-SAT problem [1–3], the simplest of an important class of optimization problems at the interface of statistical physics [4, 5] and computer science [6–8]. The XOR-SAT problem consists in finding a solution to the set of linear equations of N Boolean variables and M equations $\mathcal{A}\vec{\sigma} = \vec{\tau}$ over the Galois Field of order 2 (usually indicated as $\mathbb{GF}(2)$), where the matrix $\mathcal{A}$ is extracted from a prescribed ensemble of Boolean matrices.

The typical properties of the linear systems can be computed in the limit of large matrices and fixed density of constraints $\gamma = M/N$. For random matrices with constant row-sums (and thus Poisson-distributed column-sums), the "order parameter" γ plays a crucial role for the solution space of the corresponding random XOR-SAT problem [4]. With increasing γ, the random XOR-SAT presents three different regimes with some features of a thermodynamics phase [9]. For $\gamma < \gamma_d$ a solution can be typically found by removing iteratively all variables present in only one equation (trivial pivots in the language of Gaussian elimination [4, 10]). In this case, it can be shown that the solution space is composed of only one cluster. For $\gamma_d < \gamma < \gamma_c$ matrices have typically a non-empty "core" (the remaining part of the matrix after the recursive elimination of the trivial pivots) and finding a solution requires a number of iterations proportional to the cube of the size of the core [10]. Here, the solution space is split into many well separated clusters. Finally, for $\gamma > \gamma_c$ in the typical case solutions cannot be found (i.e. the solution space is empty).

In the field of complex networks, Boolean matrices are used to represent empirical systems with many interacting agents: each agent is labelled with an integer and the entry of the matrix $\mathcal{A}_{ij}$ is equal to one only if agent i interacts with agent j, and zero otherwise. For instance, properties of the matrix $\mathcal{A}$ are useful to control graph properties like hyperloops or critical sets of independent nodes [11]. In order to study the typical properties of such a system, it is necessary to define an ensemble of matrices which conserves characteristic properties of the empirical case. Of particular interest are matrices with a power-law distribution of column-sums, which are typical of many empirical graphs [12–15].

We have previously introduced a simple and analytically treatable Boolean random matrix ensemble with a power-law distribution of the column-sums $p(k) \sim k^{-\beta}$ and tunable β [16, 17]. This paper describes an analytical approach to the problem of the typical rank over $\mathbb{GF}(2)$ of random matrices belonging to the this ensemble and compares the results to a numerical evaluation. Previous approaches of this kind were applied to similar and more sophisticated models, but were limited to distributions of the row/column-sums with Poisson [4] or regular tails [18], or with power-law tails with a finite cut-off [10, 19].

The calculation presented here is similar to the replica calculation for spin-glasses [20]. It allows to find a formula for the typical rank in the limit of large matrices as a function of the model parameters γ and β, which allows to derive interesting phase diagrams. In particular, we estimate a second order transition in the typical rank varying the parameter γ. We compares the results with the structure of solution space obtained numerically. These results are resumed by interesting phase diagram for the behavior of the linear system with varying density of constraint γ and power-law exponent β.

2 Matrix Ensemble

This paragraph briefly describes the matrix ensemble. A more exhaustive characterization can be found in [16, 17].

The matrix ensemble (Fig. 1) was originally formulated as a null model for (biological) transcriptional regulatory networks. It is defined by the following generative algorithm. For each column of $\mathcal{A}$, (i) throw a bias from a prescribed probability distribution $\pi_M(d\theta)$ and (ii) set the column elements of $\mathcal{A}$ to be 0 or 1 according to the toss of a coin with bias θ. Since each column is thrown independently, the resulting probability law is

$$\mathfrak{p}(\mathcal{A}) = \prod_{i=1}^{N} \int_0^1 \theta_i^{\sum_{j=1}^{M} \mathcal{A}_{ij}} \, (1-\theta_i)^{\sum_{j=1}^{M}(1-\mathcal{A}_{ij})} \, \pi_M(d\theta_i). \tag{1}$$

Note that only columns are independent, while the row elements are not independent, but symmetric by permutations.

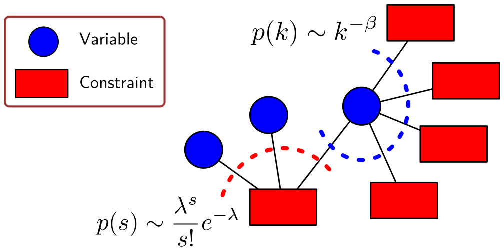

Figure 1: Schematic representation of the matrix ensemble. The probability that a variable is involved in k constraints is asymptotically proportional to a power-law $p(k) \sim k^{-\beta}$ in the limit of large matrices. Vice versa, the probability that a constraint contains s variables is a Poisson distribution $p(s) \sim \frac{\lambda^s}{s!} e^{-\lambda}$, where $\lambda > 0$ is defined in the text.

To complete the model, one has to specify the choice for $\pi_M(d\theta)$, which determines the behavior of the graph ensemble. To obtain a power-law column-sums distribution we choose the two-parameter distribution

$$\pi_M(d\theta) = Z_M^{-1} \theta^{-\beta} \chi_{\left(\frac{\alpha}{M},\ 1\right]} d\theta, \tag{2}$$

where $\alpha > 0$ and $\beta > 1$ are free parameters, $\chi_{\left(\frac{\alpha}{M},\ 1\right]}$ is the characteristic function of the interval $\left(\frac{\alpha}{M},\ 1\right]$, taking the value one inside the interval and zero everywhere else, and $Z_M = \frac{(M/\alpha)^{\beta-1}-1}{\beta-1}$ is the normalization constant. The function $\theta^{-\beta}$ of Eq. 2 gives a power-law tail to the column-sums distribution. Conversely, the cutoff on θ defined by α poses a constraint on the number of nodes with low degree, and will be used to control the probability to extract a node with small k. In the limit of large graphs (i.e. in the limit $M, N \to \infty$, with $M/N = \gamma < \infty$) the probability to extract a matrix with k_i ones in the $i-th$ column is asymptotically

$$\mathfrak{p}(\mathcal{A}) = \prod_{i=1}^{\infty} \int_0^{\infty} \frac{t_i^{k_i}\, e^{-t}}{k_i!} \pi_\infty(dt) \ , \tag{3}$$

where

$$\pi_\infty(dt) = (\beta - 1)\alpha^{\beta-1} \chi_{[\alpha,\infty)} t^{-\beta}\ dt \tag{4}$$

is the limit of the distribution in Eq. 2. Eqs. 3 and 4 imply that the probability to have a column with k ones and the probability to have a row with s ones in the limit of the large graphs are respectively $\mathfrak{p}_c(k) = \int_0^\infty \frac{t^k}{k!} e^{-t} \pi_\infty(dt) \approx k^{-\beta}$, and $\mathfrak{p}_r(s) = \frac{\lambda^s}{s!} e^{-\lambda}$, where $\lambda = \gamma \int_0^\infty t\, \pi_\infty(dt)$.

Fig. 2 reports the distribution of the nonzero entries of matrices extracted from the ensemble described by Eq. 3, for different values of α and β. As expected, the column-sums (top) follow a power-law distribution while the distribution of row-sums (bottom) follow a Poisson distribution. For $1 < \beta < 2$ the mean row-sum depends on the dimension of the system as $\mu = \frac{\beta-1}{\beta-1} \left(\frac{\alpha}{\gamma}\right)^{\beta-1} N^{2-\beta}$, while for $\beta > 2$, the mean value of the distribution is independent of the size of the system and it is $\mu = \frac{\beta-2}{\beta-1} \frac{\alpha}{\gamma}$.

3 Calculation of the Typical Rank

We will now consider the rank of a matrix belonging to the ensemble described in the previous paragraph. There are different methods for computing the rank of a given matrix $\mathcal{A}$. Here, we exploit the calculation of the number of solutions of the corresponding homogeneous linear system

$$\mathcal{N}(\mathcal{A}) = \sum_{\vec{\sigma}} \delta \left(\mathcal{A}\vec{\sigma}\right)_{(\bmod 2)},$$

where $\vec{\sigma} \in \{0,1\}^N$ and $\delta\left(\vec{\sigma}\right)_{(\mathrm{mod}\ 2)}$ is different from zero only if $\vec{\sigma} \equiv \vec{0}\,(\mathrm{mod}\,2)$. Since linear algebra applies, the number of solutions of the homogeneous system over the finite field $\mathbb{GF}(2)$ can be expressed in terms of the dimension of the kernel of matrix $\mathcal{A}$

$$\mathcal{N}(\mathcal{A}) = 2^{\mathrm{null}(\mathcal{A})}.$$

Using the rank-nullity theorem

$$\mathrm{rank}(\mathcal{A}) + \mathrm{null}(\mathcal{A}) = N,$$

the typical rank of random matrices will be

$$\langle \mathrm{rank}(\mathcal{A}) \rangle = N - \langle \log_2 \mathcal{N}(\mathcal{A}) \rangle,$$

where the average $\langle \cdot \rangle$ is carried over the matrix ensemble in Eq. 1.

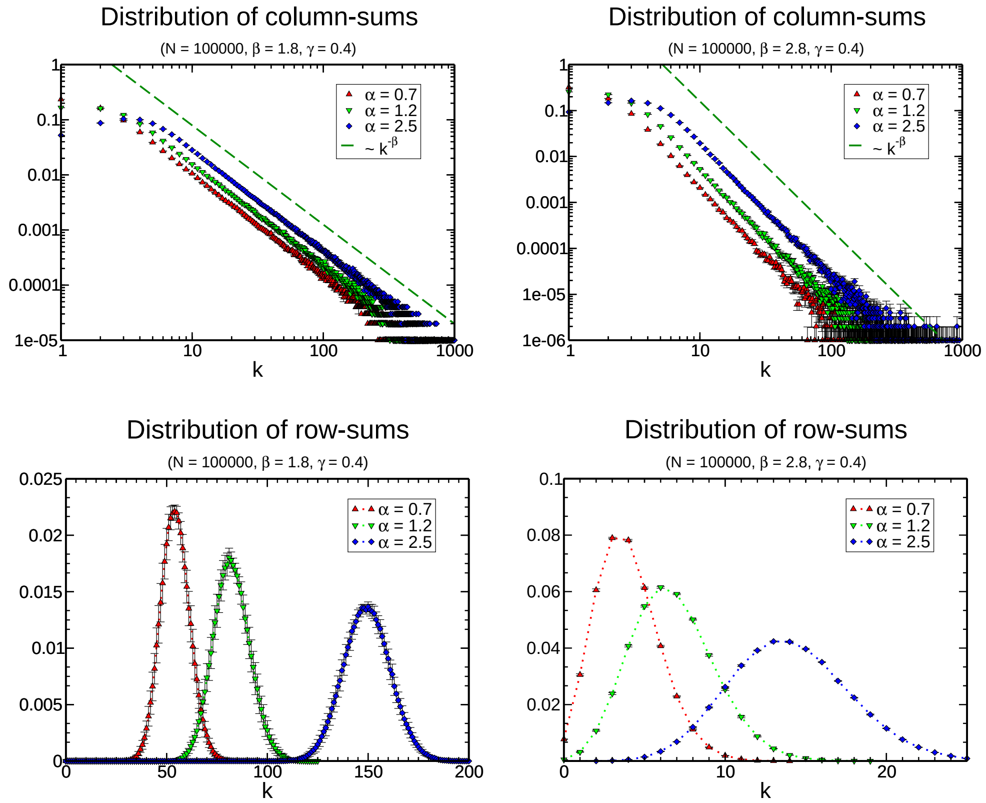

Figure 2: Distribution of nonzero entries for the matrix ensemble (Eq. 3), at $\beta = 1.8$ (left) and $\beta = 2.8$ (right). As reported in the text, the column-sums (top) follow a distribution having a power-law tail with exponent β. The dashed (green) line is a guide to the eye. On the other hand, the distribution the row-sums (bottom) follows a Poisson-like distribution with mean depending on the value of the parameter β.

In order to calculate the logarithm of the number of solutions we use the known limit

$$\log \mathcal{X} = \lim_{k\to 0} \frac{\mathcal{X}^k - 1}{k}, \tag{5}$$

where $\mathcal{X}$ is a generic random variable [20, 21]. In principle, using the above limit, it is possible to calculate the average $\langle \log \mathcal{X} \rangle$ knowing the function $\phi(k) = \left\langle \mathcal{X}^k \right\rangle$, where k is a real parameter. However, the calculation of the function $\phi(k)$ for any real k is typically hard.

As proposed in [20, 21], a feasible protocol to compute $\phi(k)$ consists in calculating the $k - th$ moment of the random variable $\left\langle \mathcal{X}^k \right\rangle$ (i.e. evaluate $\phi(k)$ for integer values of k) and then finding by interpolation a reasonable extension for any real k. In many cases [20–23] it is possible to find a well-behaved extension of the function $\phi(k)$, but this is not generally true [24].

Thus, we are interested in the calculation of the $k - th$ moment of the number of solutions $\phi(k) = \left\langle \mathcal{N}(\mathcal{A})^k \right\rangle$. As reported in Appendix A, we find

$$\left\langle \mathcal{N}_0^k(A) \right\rangle = 2^{-kM} \sum_{\{m_\mathcal{S}\}} \binom{M}{\vec{m}} \left[\sum_{\mathcal{T}\in[k]} \xi_M \left(\sum_{\mathcal{S}\in[k]} m_\mathcal{S} \,]\mathcal{S}\cap\mathcal{T}[\right) \right]^N , \tag{6}$$

where the sum is carried over 2^k integer variables labelled by an element of $[k]$ (i.e. the set of all possible subsets of $\{1, 2, \ldots, k\}$) constrained by $\sum_{\mathcal{S}\in[k]} m_\mathcal{S} = 1$ ("replica" indices), and $]\Omega[$ is equal to one if the cardinality of the set Ω is odd and zero otherwise. The function $\xi_M(h) := \int \pi_M(d\theta)\,(1-2\theta)^h$ is related to the moments of the column-sum distribution of the random matrices extracted from the ensemble in Eq. 2. Note that Eq. 6 is not an approximation but it is valid for any $M, N < \infty$.

In the limit $M \to \infty$ at fixed $x = h/M$ the function $\xi_M(h)$ can be written as

$$\xi(x) = \lim_{M\to\infty} \xi(h/M) = \int \pi_\infty(dt) e^{-x\,t},$$

where $\pi_\infty(dt) = \lim_{M\to\infty} \pi_M(d\theta)$ (see Eq. 4). It is immediate to observe that $\xi(x)$ is the moment-generating function of Eq. 1

$$\mathfrak{p}_c(k) = (-)^k k! \, \frac{d^k \xi(x)}{dx^k}.$$

Using the defining expression for the ensemble (Eq. 1), Eq. 6 can be rewritten as

$$\left\langle \mathcal{N}_0^k(\mathcal{A}) \right\rangle \approx 2^{-km} \int [dx] \exp N \left\{ \gamma \sum_{\mathcal{S} \in [k]} \mathfrak{S}(x_{\mathcal{S}}) + \right. \\ \left. + \log \left[\sum_{\mathcal{T} \in [k]} \xi_\infty \left(\sum_{\mathcal{S} \in [k]}]\mathcal{S} \cap \mathcal{T}[x_{\mathcal{S}} \right) \right] + o(1/N) \right\}, \quad (7)$$

where the integration is carried over the rescaled variables $x_{\mathcal{S}} = m_{\mathcal{S}}/N$ (with the constraint $\sum_{\mathcal{S} \in [k]} x_{\mathcal{S}} = 1$) and $\sum_{\mathcal{S} \in [k]} \mathfrak{S}(x_{\mathcal{S}}) = \sum_{\mathcal{S} \in [k]} -x_{\mathcal{S}} \log x_{\mathcal{S}}$ is the Shannon entropy. The above expression diverges exponentially with the dimension N of the matrices, and thus it is possible to use the saddle point approximation. In order to compute the saddle point, it is necessary to find the maximum of Eq. 7 varying $x_{\mathcal{S}}$, i.e. it is necessary to solve a system of a 2^k variables for any integer k. Obviously, this is unfeasible and one must impose a symmetry ansatz for the saddle point solution in order to reduce the number of variables.

The simplest hypothesis it that the most symmetric solution would dominate (in the theory of glassy systems this solution is usually called replica symmetric (RS) solution)

$$\begin{aligned} &x_\emptyset = x \\ &x_{\mathcal{S}} = 1 - (2^k - 1)x, \quad \mathcal{S} \neq \emptyset, \end{aligned}$$

where all variables are equal, except one in order to satisfy the constraint $\sum_{\mathcal{S} \in [k]} x_{\mathcal{S}} = 1$. Here, the variable x plays the same role of the "Edward-Anderson" order parameter in the Spin Glass theory [20]: for $x = 0$, the total entropy $\sum_{\mathcal{S} \in [k]} \mathfrak{S}(x_{\mathcal{S}})$ is exactly zero and then only one state, i.e. the most symmetric state, dominates the saddle point in Eq. 7. On the contrary, for $x = 1$, the total entropy assumes the highest possible value and then many different states contribute to the saddle point in Eq. 7.

Using the RS ansatz, the asymptotic behaviour of the $k - th$ moment of the number of solutions can be written as

$$\frac{\log \overline{\mathcal{N}_0^k(\mathcal{A})}}{N} = -k\gamma \log 2 + \max_x \Big\{ \gamma(2^k - 1)\mathfrak{S}(x) + \\ + \gamma\mathfrak{S}(1-x) + \log\left[1 + (2^k - 1)\xi(2^{k-1}x)\right] \Big\}. \quad (8)$$

It is important to observe that the variable k in Eq. 8 can assume any real value and it can be considered as a possible extension of the Eq. 6 in the limit of large matrices. Eq. 8 depends directly on the chosen symmetry ansatz and is not guaranteed to be consistent. In our case, we will show that Eq. 8 gives results that agree with numerical results.

We can now take the limit $k \to 0$. Thus we have

$$\frac{\langle rank(\mathcal{A})\rangle}{N} = 1 - \lim_{N\to\infty} \frac{\langle \log_2 \mathcal{N}_0 \rangle}{N} = \max_{x\in[0,1]} \left\{ \gamma\mathfrak{S}_0(x) - \gamma + \xi\left(\frac{x}{2}\right)\right\}, \quad (9)$$

where $\mathfrak{S}_0(x) = -x\log x + x$. The above equation can be used directly to find the typical rank of the matrices extracted from the matrix ensemble proposed in Eq. 1. Fig. 3 compares the theoretical prediction of the typical rank with simulations. It is possible to observe that, independently of the choice of the parameters α and β, the theoretical prediction is in good agreement with the simulations.

Interestingly, the theoretical prediction of the rank (Eq. 9) can have a second order discontinuity varying the density of constraints γ, due to the fact that the value of the RS order parameter x which maximize the expression in Eq. 9 can have a jump (Fig. 4). In particular, we find that for any $\beta > 2$ there exists a critical value $\alpha_c(\beta)$ such as for $\alpha < \alpha_c(\beta)$ there are no jumps varying the parameter γ. Instead, for $\alpha > \alpha_c(\beta)$, it is possible to identify a critical value $\gamma_c(\beta)$ in which x has a jump. On the contrary, for $1 < \beta < 2$ a discontinuity is always present.

The presence of a second order discontinuity of the typical rank is a signal of the fact that the totally symmetric solution (RS solution) is no longer valid (even if it may still be a good approximation for the calculation of the typical rank) caused by a spontaneous symmetry breaking of the solution space in many well-separated clusters [20]. In this case, a less symmetric solution (called replica symmetry breaking (RSB) solution) dominates the saddle point in Eq. 7. We did not explore analytically this regime.

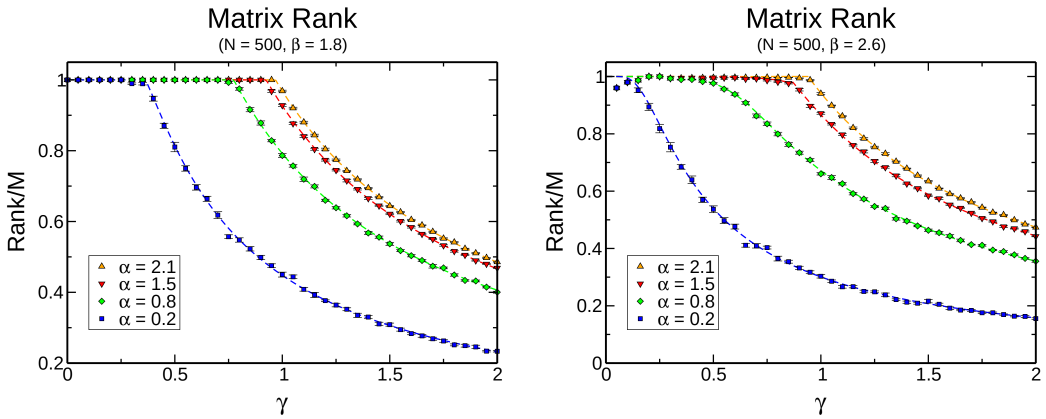

Figure 3: Distribution of the typical rank obtained from simulation with $N = 500$ and $\beta = 1.8$ (left) or $\beta = 2.6$ (right), varying the parameter γ. As shown in the figures, the numerical data are in agreement with the theoretical prediction obtained by Eq. 9. The deviation for small values of γ is due to the small system size.

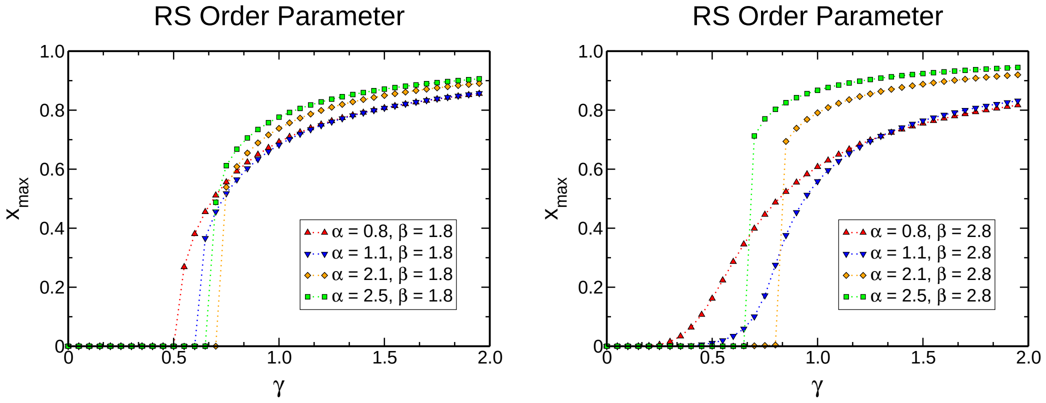

Figure 4: Value of the RS order parameter x that maximizes Eq. 9 at fixed β. For $\beta < 2$ (left), for any value of α there exists a critical value of γ in which the value of x at the maximum has a jump. For $\beta \geq 2$ (right) and α sufficiently small, the value x_{max} does not have any discontinuity. Otherwise, it is possible to identify a γ_c (that depends on α and β) for which the value of x_{max} has a jump.

$$\begin{array}{ll}(\alpha) & x_2 + x_3 \equiv 0, \ \bmod\ 2 \\ (\beta) & x_1 + x_2 + x_3 \equiv 0, \ \bmod\ 2 \\ (\gamma) & x_1 + x_3 + x_4 \equiv 0, \ \bmod\ 2\end{array}$$

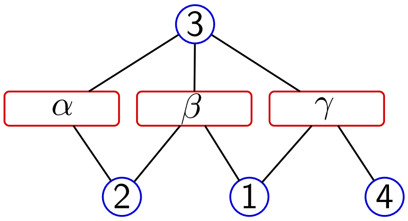

Figure 5: Factor graph representation of the XOR-SAT problem. In the sketch the variables (columns) are represented by circles and the constraints (rows) by rectangles.

4 Leaf Removal and Organization of the Solution Space

As described in the previous paragraph, the typical rank of $\mathcal{A}$ is related to the total number of solutions of the linear system $\mathcal{A}\vec{\sigma} = \vec{0}$. In particular, we found an analytical expression for the typical rank which has sharp transitions when the parameters that define the matrix ensemble vary continuously. As previously discussed, these transitions are related to the clusterization of the solution space. This paragraph focuses on the geometrical organization of the solution space of the linear system $\mathcal{A}\vec{\sigma} = \vec{\tau}$ (the XOR-SAT problem) and the comparison between numerical evaluations and our theoretical predictions. A general introduction to this problem can be found in [4,10,25].

A system of linear equations in $\mathbb{GF}(2)$ can be conveniently represented by factor graphs, defined by the matrix $\mathcal{A}$, in which variables and constraints correspond to distinct types of nodes. If the variable i is present in the constraint α, a link (i, α) is drawn in the factor graph (Fig. 5).

Following [4,10,25], it is possible to obtain a precise definition of clusters of solutions using the so-called "leaf removal" algorithm. The leaf removal algorithm is an iterative algorithm used to gradually eliminate all trivially constrained variables (called trivial pivots in the language of Gaussian elimination). It is easy to prove that when a variable (called "leaf") is connected to only one constraint, it is always possible to choose its value such that the constraint is always satisfied (e.g. variable 4 in Fig. 5). The leaf removal algorithm is based on this evidence and it is defined as follows: (i) pick a variable that appears only in one constraint (leaf) and (ii) remove it together with the only constraint it is connected to. The process is iterated until no leaves remain. The part of the factor graph that cannot be removed by leaf removal iteration is called “core” and does not depend on the order in which the leafs are removed. In this case, the order parameter of the reduced linear system will be

$$\gamma_{core} = \frac{M_{core}}{N_{core}}, \tag{10}$$

i.e. the density of constraints that are not trivially satisfied.

The presence of the core is related to the clusterization of the solution space. If $\gamma_{core} = 0$ (no core is present), the problem to find a solution of the linear system $\mathcal{A}\vec{\sigma} = \vec{\tau}$ is trivial (the complete solution can be found by running the leaf removal in reverse direction, in a scheme usually called leaf reconstruction) and the solution space is composed of only one cluster. If $0 < \gamma_{core} < 1$, the core is not trivial (but not over-constrained) and each solution of the linear system reduced to the core variable defines a single cluster. All the solutions built from a core solution by leaf reconstruction belong to the same cluster. Finally, for $\gamma_{core} > 1$ the reduced linear system for the core variables is over-constrained, so that no solutions are typically found.

Fig. 6 reports the curves of the typical γ_{core} varying the density of constraints γ obtained by numerical simulations of the leaf removal algorithm. As predicted in the previous paragraph, the presence of a non over-constrained core depends on the choice of the parameter β. For $\beta < 2$ (left panel), varying the parameter γ it is always possible to identify three regimes: an empty core phase ($\gamma_{core} = 0$), a non over-constrained core phase ($\gamma_{core} < 1$) and an over-constrained core phase ($\gamma_{core} > 1$). On the other hand, for $\beta > 2$ (right panel) the not over-constrained ($\gamma_{core} < 1$) core is present only for α sufficiently large. All these results are resumed in the phase diagrams obtained from in Fig. 7.

5 Conclusion

In conclusion, we have presented a simple calculation of the typical rank of random matrices with power-law distributed column-sums on the Galois Field of order 2. The matrices can describe a graph or a sparse linear system for Boolean variables. The calculation is based on a fairly standard replica-like approach, where we compute the generic k-th moment of the number of

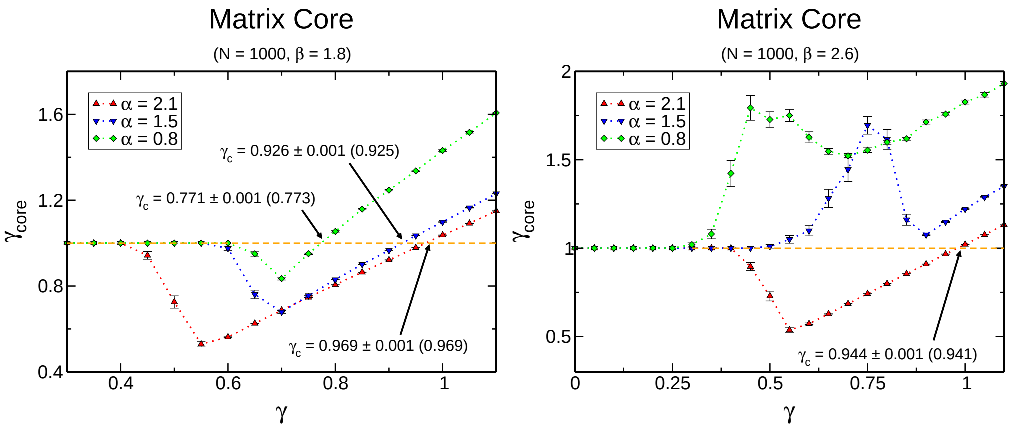

Figure 6: Numerical simulation of the γ_{core} varying the parameter γ, for different value of α and β. For $\beta < 2$ (left), it is always possible to find the critical value γ_c of inversion of the core. For $\beta > 2$ (right), only for α sufficiently large it is possible to find the critical value γ_c. In parenthesis the theoretical predictions.

solutions of the associated linear system and we consider the limit $k \to 0$ of its analytical extension in the maximally symmetric case.

Differently from other models present in the literature [4, 10, 18, 19], the simplicity of the matrix ensemble [16] that we employ here allows to find an analytical expression for the typical rank without having to impose any cutoff on the power-law distribution. As shown in Figs. 3, the typical rank calculated with our method is in fairly good agreement with the numerical results. We find that, as usually happens in this kind of models [4,10] the typical rank can have a second order discontinuity with increasing density of constraints γ. This discontinuity is related to the clusterization of the solution space in many well separated clusters of the related XOR-SAT problem [25]. Our result indicates that the same phenomenology can exist in presence of truly power-law tails.

More in detail, since the matrix ensemble is defined as a function of the model parameters α, which sets a lower cutoff on the row-sums and β, the exponent of the column-sum distribution, one can study the variation of this threshold with "phase diagrams" where these parameters vary together with the density of constraints. Specifically, the presence of the typical rank discontinuity at $\gamma = \gamma_c$ depends on the choice of α and β. For $\beta < 2$ the

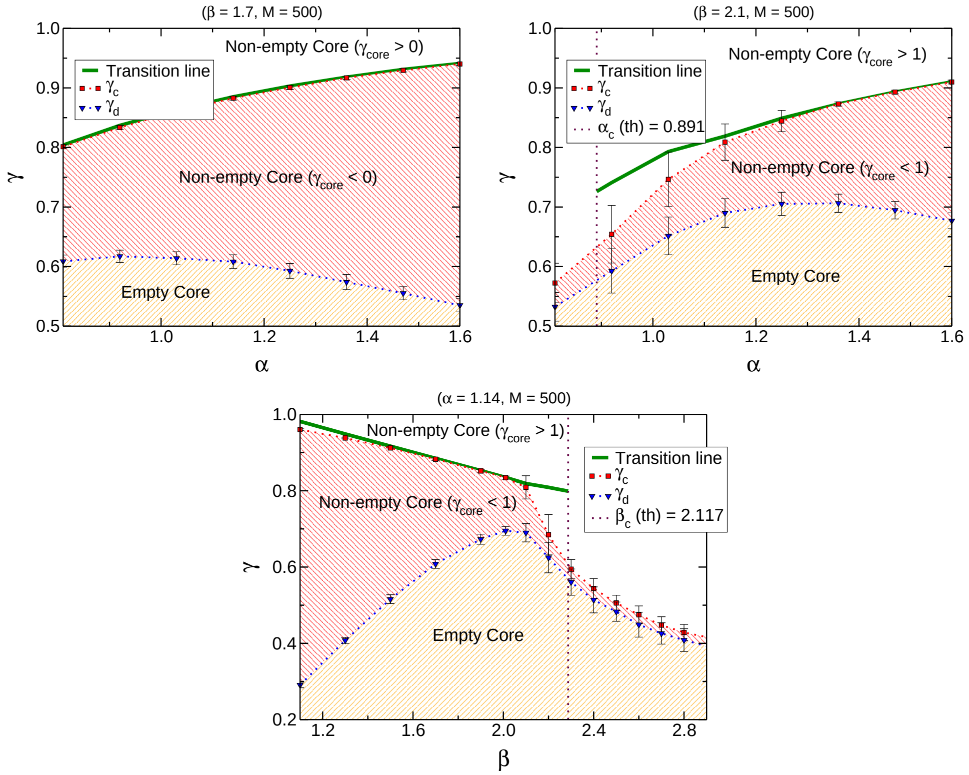

Figure 7: Phase diagrams obtained from simulation with $M = 500$ and $N = M/\gamma$. In the plots the results of the simulation are compared with the theoretical predictions (solid lines). The top panels contain fixed β phase diagrams for $\beta > 2$ and $\beta < 2$, while the bottom panel is a fixed α phase diagram. The largest errors in the phase diagrams arise around the critical value α_c (top panels) and β_c (bottom panel). This can be explained observing that near such critical values the matrix core is very small compared to the finite size fluctuations.

discontinuity exists for any choice of α. Otherwise, for $\beta > 2$, it is possible to identify a critical value $\alpha_c(\beta)$ such that only for $\alpha > \alpha_c(\beta)$ a critical value γ_c exists.

The role played by α in this model is similar to that played by the constraint connectivity K in the K–XOR-SAT problem [2]. In this case, the row-sum of the matrix is equal to K, and the clustering of solution is possible only if $K > 2$. In our case, it is simple to verify that only for $\beta < 2$ the fraction of rows with two nonzero entries always vanishes for every α in the large N limit. Thus, we speculate that the density of rows with two or less nonzero entries may become important and affect in some cases the phase diagram for $\beta > 2$, causing the observed lack of the clusterization regime.

Finally, the approach presented here, suitably generalized, may be useful to study self-organizing properties of systems with many interacting agents, where similar threshold phenomena can emerge as a function of the properties of the network that defines the agent interactions. In this case, the parameters of the matrix ensemble represent tunable quantitative topological properties of the interaction network such as the connectivity and the density of interactions.

A Calculation of $\left\langle \mathcal{N}_0^k(\mathcal{A}) \right\rangle$

In this appendix we explicitly calculate the $k-th$ moment of the number of solutions of the homogeneous linear system $\mathcal{N}_0(\mathcal{A}) = \sum_{\vec{\sigma}} \delta\left(\mathcal{A}\vec{\sigma}\right)$, with $\vec{\sigma} \in \{0,1\}^N$ and $\delta\left(\vec{\sigma}\right) = 1$ only if $\vec{\sigma} = 0$. Let $\mathfrak{p}(\mathcal{A})$ a generic probability distribution for the random matrix $\mathcal{A}$: thus the $k-th$ momentum can be written as

$$\left\langle \mathcal{N}_0^k(\mathcal{A}) \right\rangle = \sum_{\mathbb{X} \in \{0,1\}^N \otimes \{0,1\}^k} \; \sum_{\mathcal{A} \in \{0,1\}^M \otimes \{0,1\}^N} \mathfrak{p}(\mathcal{A}) \prod_{j=1}^{M} \prod_{\alpha=1}^{k} \delta\left(\sum_{i=1}^{N} \mathcal{A}_{ji} \mathbb{X}_{i\alpha} \right).$$

For simplicity, in the rest of the appendix we use the convention

$$\begin{aligned} &i \in \mathbb{N},\, i = 1,\, \ldots,\, N \text{ (position of the row)} \\ &j \in \mathbb{N},\, i = 1,\, \ldots,\, M \text{ (position of the column)} \\ &\alpha \in \mathbb{N},\, i = 1,\, \ldots,\, k \text{ (number of the ``replica'')} \end{aligned}$$

Use probability distribution for our model (Eq. 1), the expression of the $k-th$ moment will be

$$\begin{aligned}\left\langle \mathcal{N}_0^k(\mathcal{A})\right\rangle &= \sum_{\mathbb{X}}\sum_{A}\mathfrak{p}(\mathcal{A})\prod_{j,\alpha}\frac{1+(-1)^{\sum_i \mathcal{A}_{ji}\mathbb{X}_{i\alpha}}}{2} = \\ &= 2^{-kM}\sum_{\mathbb{X}}\sum_{\mathcal{A}}\int\left[\prod_i \pi_M(d\theta_i)\right]\cdot \\ &\cdot\left[\prod_{j,\alpha}\left(1+(-1)^{\sum_i \mathcal{A}_{ji}\mathbb{X}_{i\alpha}}\right)\right]\left[\prod_j \theta_i^{\sum_i \mathcal{A}_{ji}}\left(1-\theta_i\right)^{M-\sum_i \mathcal{A}_{ji}}\right],\end{aligned}$$

where we used the explicit representation of the Kronecker delta for binary variables

$$\delta\left(\sigma\right) = \frac{1+(-1)^{\sigma}}{2}.$$

At this level, it is possible to exchange the sums over $\mathcal{A}$ and the integration to obtain

$$\begin{aligned}\left\langle \mathcal{N}_0^k(\mathcal{A})\right\rangle &= 2^{-kM}\sum_{\mathbb{X}}\int\left[\prod_i \pi_M(d\theta_i)\right]\cdot \\ &\cdot\prod_j\left[\sum_{\vec{a}\in\{0,1\}^N}\prod_\alpha\left(1+(-1)^{\sum_i a_i\mathbb{X}_{i\alpha}}\right)\prod_i \theta_i^{a_i}\left(1-\theta_i\right)^{1-a_i}\right].\end{aligned}$$

The last term does not depend explicitly on j and then we above expression can be rewritten as

$$\begin{aligned}\left\langle \mathcal{N}_0^k(\mathcal{A})\right\rangle &= 2^{-kM}\sum_{\mathbb{X}}\int\left[\prod_i \pi_M(d\theta_i)\right]\cdot \\ &\cdot\left[\sum_{\vec{a}\in\{0,1\}^N}\prod_\alpha\left(1+(-1)^{\sum_i a_i\mathbb{X}_{i\alpha}}\right)\prod_i \theta_i^{a_i}\left(1-\theta_i\right)^{1-a_i}\right]^M.\end{aligned}$$

Now, using the identity

$$\prod_{\alpha=1}^{k}\left(1+f(\alpha)\right) = \sum_{\mathcal{S}\subseteq[k]}\prod_{\alpha\in\mathcal{S}} f(\alpha),$$

where $[k]$ is the set of all the possible subsets of $\{1, \ldots, k\}$, the above expression becomes

$$\left\langle \mathcal{N}_0^k(\mathcal{A}) \right\rangle = 2^{-kM} \sum_{\mathbb{X}} \int \left[\prod_i \pi_M(d\theta_i) \right] \cdot \\ \cdot \left\{ \sum_{\mathcal{S} \subseteq [k]} \prod_i \left[\sum_{\sigma \in \{0,1\}} \left((-1)^{\sum_{\alpha \in \mathcal{S}} \mathbb{X}_{i\alpha}} \theta_i \right)^{\sigma} (1-\theta_i)^{1-\sigma} \right] \right\}^M .$$

It easy to observe that the last term can be directly calculated. Thus, after a sum over σ we obtain

$$\sum_{\sigma \in \{0,1\}} \left((-1)^{\sum_{\alpha \in \mathcal{S}} \mathbb{X}_{i\alpha}} \theta_i \right)^{\sigma} (1-\theta_i)^{1-\sigma} = 1 - 2\theta_i \, \delta \left(1, \sum_{\alpha \in \mathcal{S}} \mathbb{X}_{i\alpha} \right),$$

where $\delta(1, \sigma)$ equals 1 if and only if $\sigma = 1$, and

$$\left\langle \mathcal{N}_0^k(\mathcal{A}) \right\rangle = 2^{-kM} \sum_{\mathbb{X}} \int \left[\prod_i \pi_M(d\theta_i) \right] \left\{ \sum_{\mathcal{S} \subseteq [k]} \prod_i \left[1 - 2\theta_i \, \delta \left(1, \sum_{\alpha \in \mathcal{S}} \mathbb{X}_{i\alpha} \right) \right] \right\}^M .$$

In order to complete the calculation, it is necessary to expand the term inside the curly brackets. Let $\{m_{\mathcal{S}}\}$ the set of 2^k variables such that $\sum_{\mathcal{S} \in [k]} m_{\mathcal{S}} = M$. Thus we have

$$\left\langle \mathcal{N}_0^k(\mathcal{A}) \right\rangle = 2^{-kM} \sum_{\mathbb{X}} \sum_{\{m_{\mathcal{S}}\}} \binom{M}{\vec{m}} \cdot \\ \cdot \prod_i \left\{ \int \pi_M(d\theta_i) \prod_{\mathcal{S} \subseteq [k]} \left[1 - 2\theta_i \, \delta \left(1, \sum_{\alpha \in \mathcal{S}} \mathbb{X}_{i\alpha} \right) \right]^{m_{\mathcal{S}}} \right\},$$

where $\binom{M}{\vec{m}}$ is the multinomial. Using the simple identity

$$\left[1 - 2\theta_i \, \delta \left(\sum_{\alpha \in \mathcal{S}} \mathbb{X}_{i\alpha} \right) \right]^{m_{\mathcal{S}}} = (1 - 2\theta_i)^{\delta\left(\sum_{\alpha \in \mathcal{S}} \mathbb{X}_{i\alpha} \right) m_{\mathcal{S}}} ,$$

we obtain

$$\left\langle \mathcal{N}_0^k(A) \right\rangle = 2^{-kM} \sum_{\{m_{\mathcal{S}}\}} \binom{M}{\vec{m}} \prod_i \left\{ \sum_{\mathbb{X}_i \in \{0,1\}^k} \xi_M \left(\sum_{\mathcal{S} \subseteq [k]} \delta \left(\tilde{\mathbb{X}}_i(\mathcal{S}) \right) m_{\mathcal{S}} \right) \right\},$$

where we used the notation

$$\tilde{\mathbb{X}}_i(\mathcal{S}) := \sum_{\alpha\in\mathcal{S}} \mathbb{X}_{i\alpha}$$
$$\xi_M(h) := \int \pi_M(d\theta)\,(1-2\theta)^h\,.$$

It is immediate to observe that the expression inside the curly bracket is independent on i:

$$\left\langle \mathcal{N}_0^k(A) \right\rangle = 2^{-kM} \sum_{\{m_\mathcal{S}\}} \binom{M}{\vec{m}} \left[\sum_{\vec{x}\in\{0,1\}^k} \xi_M \left(\sum_{\mathcal{S}\subseteq[k]} \delta\left(\tilde{x}(\mathcal{S})\right) m_\mathcal{S} \right) \right]^N , \qquad (11)$$

where $\tilde{x}(\mathcal{S}) = \sum_{\alpha\in\mathcal{S}} \vec{x}_\alpha$. The above expression can be simplified if we define $\mathcal{T}$ as the set of the positions of the vector $\vec{x}$ different from zero. Indeed, the function $\tilde{x}(\mathcal{S})$ can be expressed as

$$\tilde{x}(\mathcal{S}) =]\mathcal{S}\cap\mathcal{T}[$$

where $]\Omega[\; = 1$ if the cardinality of Ω is odd and zero otherwise. Thus, replacing the sum over $\vec{x}$ with the sum over $\sum_{\mathcal{T}\subseteq[k]}$ in Eq. 11 we finally obtain

$$\left\langle \mathcal{N}_0^k(A) \right\rangle = 2^{-kM} \sum_{\{m_\mathcal{S}\}} \binom{M}{\vec{m}} \left[\sum_{\mathcal{T}\subseteq[k]} \xi_M \left(\sum_{\mathcal{S}\subseteq[k]} m_\mathcal{S}]\mathcal{S}\cap\mathcal{T}[\right) \right]^N .$$

References

[1] O. Dubois, J. Mandler. The 3-XORSAT threshold. *Comptes Rendus Mathematique*, 335(11):963–966, 2002.

[2] R. Monasson. Introduction to Phase Transitions in Random Optimization Problems. *Lecture Notes of the Les Houches Summer School on Complex Systems, Elsevier*, 2007.

[3] M. Mézard, A. Montanari. *Information, physics, and computation.* Oxford University Press, USA, 2009.

[4] M. Mézard, F. Ricci-Tersenghi, R. Zecchina. Two solutions to diluted p-spin models and XORSAT problems. *Journal of Statistical Physics*, 111(3):505–533, 2003.

[5] A. Montanari, F. Ricci-Tersenghi. On the nature of the low-temperature phase in discontinuous mean-field spin glasses. *The European Physical Journal B-Condensed Matter and Complex Systems*, 33(3):339–346, 2003.

[6] P. Cheeseman, B. Kanefsky, W.M. Taylor. Where the really hard problems are. In *Proceedings of the 12th IJCAI*, pages 331–337. Citeseer, 1991.

[7] B. Selman, H. Levesque, D. Mitchell. A new method for solving hard satisfiability problems. In *Proceedings of the tenth national conference on artificial intelligence*, pages 440–446. Citeseer, 1992.

[8] D. Mitchell, B. Selman, H. Levesque. Hard and easy distributions of SAT problems. In *Proceedings of the National Conference on Artificial Intelligence*, pages 459–459. Citeseer, 1992.

[9] F. Altarelli, R. Monasson, F. Zamponi. Relationship between clustering and algorithmic phase transitions in the random k-XORSAT model and its NP-complete extensions. In *Journal of Physics: Conference Series*, volume 95, page 012013. IOP Publishing, 2008.

[10] A. Braunstein, M. Leone, F. Ricci-Tersenghi, R. Zecchina. Complexity transitions in global algorithms for sparse linear systems over finite fields. *Journal of Physics A: Mathematical and General*, 35:7559, 2002.

[11] V.F. Kolchin. *Random graphs*. Cambridge Univ Pr, 1999.

[12] R. Albert, H. Jeong, A.L. Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

[13] A.L. Barabási, R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[14] H. Jeong, B. Tombor, R. Albert, Z.N Oltvai, A.L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[15] N. Guelzim, S. Bottani, P. Bourgine, F. Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nature genetics*, 31(1):60–63, 2002.

[16] F. Bassetti, M. Cosentino Lagomarsino, B. Bassetti, P. Jona. Random networks tossing biased coins. *Physical Review E*, 75(5):56109, 2007.

[17] F. Bassetti, M. Cosentino Lagomarsino, S. Mandrà. Exchangeable random networks. *Internet Mathematics*, 4(4):357–400, 2007.

[18] S. Franz, M. Leone, F.L. Toninelli. Replica bounds for diluted non-Poissonian spin systems. *Journal of Physics A: Mathematical and General*, 36:10967, 2003.

[19] R.C. Alamino, D. Saad. Typical kernel size and number of sparse random matrices over Galois fields: A statistical physics approach. *Physical Review E*, 77(6):61123, 2008.

[20] M. Mézard, G. Parisi, M.A. Virasoro. *Spin glass theory and beyond.* World scientific Singapore, 1987.

[21] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613–2626, 1981.

[22] S. Franz, M. Leone, F. Ricci-Tersenghi, R. Zecchina. Exact solutions for diluted spin glasses and optimization problems. *Physical Review Letters*, 87(12):127209, 2001.

[23] R. Oppermann, D. Sherrington. Scaling and Renormalization Group in Replica-Symmetry-Breaking Space: Evidence for a Simple Analytical Solution of the Sherrington-Kirkpatrick Model at Zero Temperature. *Physical review letters*, 95(19):197203, 2005.

[24] J.J.M. Verbaarschot, M.R. Zirnbauer. Critique of the replica trick. *Journal of Physics A: Mathematical and General*, 18:1093, 1985.

[25] T. Mora, M. Mézard. Geometrical organization of solutions to random linear Boolean equations. *Journal of Statistical Mechanics: Theory and Experiment*, 2006:P10007, 2006.